Regular Hilberg Processes: An Example of Processes with a Vanishing Entropy Rate

Łukasz Dębowski*

Abstract

A regular Hilberg process is a stationary process that satisfies both a hyperlogarithmic growth of maximal repetition and a power-law growth of topological entropy, which are a kind of dual conditions. The hyperlogarithmic growth of maximal repetition has been experimentally observed for texts in natural language, whereas the power-law growth of topological entropy implies a vanishing Shannon entropy rate and thus probably does not hold for natural language. In this paper, we provide a constructive example of regular Hilberg processes, which we call random hierarchical association (RHA) processes. Our construction does not apply the standard cutting and stacking method. For the constructed RHA processes, we demonstrate that the expected length of any uniquely decodable code is orders of magnitude larger than the Shannon block entropy of the ergodic component of the RHA process. Our proposition supplements the classical result by Shields concerning nonexistence of universal redundancy rates.

Keywords: maximal repetition, topological entropy, entropy rate, asymptotically mean stationary processes

*Ł. Dębowski is with the Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland (e-mail: ldebowsk@ipipan.waw.pl).

An extended abstract of this paper, entitled "Regular Hilberg Processes: Nonexistence of Universal Redundancy Ratios", was presented at the 8th Workshop on Information Theoretic Methods in Science and Engineering, Copenhagen, June 24-26, 2015.

I Main ideas and results

Throughout this paper we identify stationary processes with their distributions (stationary measures) and we use terms “measure” and “process” interchangeably. Consider thus a stationary measure μ on the measurable space of infinite sequences $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ from a finite alphabet $\mathbb{A} \subset \mathbb{N}$. The random symbols will be denoted as $\xi_i : \mathbb{A}^{\mathbb{N}} \ni (x_i)_{i \in \mathbb{N}} \mapsto x_i \in \mathbb{A}$, whereas blocks of symbols will be denoted as $x_{k:l} = (x_i)_{i=k}^{l}$. The expectation with respect to μ is denoted as $\mathbf{E}_\mu$. We also use shorthand $\mu(x_{1:m}) = \mu(\xi_{1:m} = x_{1:m})$. The Shannon block entropy of measure μ is function

$$H_\mu(m) := \mathbf{E}_\mu \left[-\log \mu(\xi_{1:m}) \right], \tag{1}$$

and the Shannon entropy rate of μ is the limit

$$h_\mu := \inf_{m \in \mathbb{N}} \frac{H_\mu(m)}{m} = \lim_{m \to \infty} \frac{H_\mu(m)}{m}. \tag{2}$$

Let us introduce two functions of an individual block $\xi_{1:k}$. The first one is the maximal repetition

$$L(\xi_{1:k}) := \max \left\{ m : \text{some } x_{1:m} \text{ is repeated in } \xi_{1:k} \right\} \tag{3}$$

[1, 2, 3, 4, 5], whereas the second one is the topological entropy

$$H_{top}(m|\xi_{1:k}) := \log \operatorname{card} \left\{ x_{1:m} : x_{1:m} \text{ is a substring of } \xi_{1:k} \right\}, \tag{4}$$

which is the logarithm of subword complexity [6, 7, 1, 8, 9, 10]. In this paper, we are interested in the following class of stationary processes, defined using the Big O notation:

Definition 1 (a variation of a definition in [11]) *A stationary measure μ on the measurable space of infinite sequences $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ is called a* regular Hilberg process with an exponent $\beta \in (0, 1)$ *if it satisfies conditions*

$$L(\xi_{1:m}) = \Theta\left((\log m)^{1/\beta} \right), \tag{5}$$

$$H_{top}(m|\xi_{1:\infty}) = \Theta\left(m^{\beta} \right). \tag{6}$$

μ-almost surely, where the lower bound for the maximal repetition and the upper bound for the topological entropy are uniform in $\xi_{1:\infty}$.

The original definition in [11] uses condition $H_\mu(m) = \Theta\left(m^\beta\right)$ rather than (6) and condition $\mathbf{E}_\mu L(\xi_{1:m}) = \Theta\left((\log m)^{1/\beta}\right)$ instead of (5). Condition $H_\mu(m) = \Theta\left(m^\beta\right)$ has been originally contemplated by Hilberg [12], hence follows the name of the class of processes. Conditions (5) and (6) are, however, more natural since they pertain to an individual sequence $\xi_{1:\infty}$ and are dual in view of the following proposition:

Theorem 1 ([13]) *If $H_{top}(m|\xi_{1:k}) < \log(k - m + 1)$ then $L(\xi_{1:k}) \geq m$.*

Proof: String $\xi_{1:k}$ contains $k - m + 1$ substrings of length m (on overlapping positions). Among them there can be at most $\exp(H_{top}(m|\xi_{1:k}))$ different substrings. Since $\exp(H_{top}(m|\xi_{1:k})) < k - m + 1$, there must be some repeat of length m. Hence $L(\xi_{1:k}) \geq m$. □

In particular, since $H_{top}(m|\xi_{1:k}) \le H_{top}(m|\xi_{1:\infty})$, Theorem 1 yields

$$\begin{aligned} H_{top}(m|\xi_{1:\infty}) = O\left(m^{\beta}\right) &\Rightarrow L(\xi_{1:m}) = \Omega\left((\log m)^{1/\beta}\right), \\ L(\xi_{1:m}) = O\left((\log m)^{1/\beta}\right) &\Rightarrow H_{top}(m|\xi_{1:\infty}) = \Omega\left(m^{\beta}\right). \end{aligned}$$

Now we can see that the lower bound in (5) is implied by the upper bound in (6), whereas the upper bound in (5) implies the lower bound in (6). We might therefore suppose that conditions (5) and (6) hold simultaneously indeed for some class of processes.

Why is this problem important? In fact, according to some experimental measurements of maximal repetition, the hyperlogarithmic growth (5) holds approximately with $\beta \approx 0.4$ for texts in English, French, and German, where the lower bound for the growth of maximal repetition seems uniform, i.e., text-independent [14, 13]. Thus understanding how to construct some class of processes satisfying condition (5) may contribute to an improvement in statistical models of natural language. Although condition $H_\mu(m) = \Theta\left(m^{\beta}\right)$, related to (6), was actually considered in [12] as a hypothesis for natural language, here we should admit that the combination of conditions (5) and (6) is likely too strong to be required from the natural language models. As we will show, the power law (6) implies a vanishing Shannon entropy rate, $h_\mu = 0$, whereas the overwhelming empirical evidence asserts that the Shannon entropy rate of natural language is strictly positive, about 1 bit per character [15, 16, 17, 18, 19, 20]. Nevertheless, constructing stationary processes that satisfy the hyperlogarithmic growth (5) is nontrivial enough, so it may be illuminating to consider first a somewhat unrealistic class of processes that also satisfy the power law (6).

For the regular Hilberg processes there are two general results. As mentioned, it can be seen easily that the power law (6) implies a vanishing Shannon entropy rate.

Theorem 2 *We have $h_\mu = 0$ for a regular Hilberg process μ.*

Proof: The argument involves the random ergodic measure $F = \mu(\cdot|\mathcal{I})$, where $\mathcal{I}$ is the shift-invariant algebra [21, 22]. By the ergodic theorem for stationary processes [21], we have μ-almost surely

$$H_{top}(m|\xi_{1:\infty}) \ge \log \operatorname{card} \{x_{1:m} : F(x_{1:m}) > 0\} \ge H_F(m), \tag{7}$$

so $h_F = 0$ follows from (6), whereas as shown in [22, 23] we have

$$h_\mu = \mathbf{E}_\mu \, h_F, \tag{8}$$

from which $h_\mu = 0$ follows. □

Moreover, the ergodic decomposition of a regular Hilberg process, as defined in Definition 1, consists of ergodic regular Hilberg processes. Namely, we have:

Theorem 3 *For a regular Hilberg process μ with exponent β, the random ergodic measure $F = \mu(\cdot|\mathcal{I})$, where $\mathcal{I}$ is the shift-invariant algebra, μ-almost surely constitutes an ergodic regular Hilberg process with exponent β.*

Proof: We have $\mu = \int F d\mu$. Hence every event of full measure μ must be μ-almost surely an event of full measure F. This implies the claim. □

We suppose that the above property is not true for the original definition of a regular Hilberg process given in article [11], but we do not investigate this problem in this paper.

We will present now some constructive example of regular Hilberg processes. The example will be called random hierarchical association (RHA) processes. The RHA processes are parameterized by certain free parameters which we will call perplexities (a name borrowed from computational linguistics). Approximately, perplexity k_n is the number of distinct blocks of length 2^n that appear in the process realization. Exactly in this meaning, term "perplexity" is used in computational linguistics. It turns out that controlling perplexities, we can control the value of the Shannon block entropy and force the Shannon entropy rate to be zero. It turns out as well that we can control the value of the topological entropy and the maximal repetition. In this way we can construct a stationary process exhibiting quite an arbitrary desired growth of the topological entropy and the maximal repetition, such a regular Hilberg process.

We have invented the RHA processes as a construction unrelated to the cutting and stacking method [24], used for constructing stationary processes with certain desired properties. The cutting and stacking method seems more abstract and more general than the RHA process method. Certainly, these two methods adopt very different strategies. The cutting and stacking method, being a tool borrowed from ergodic theory, approximates the constructed process by an abstract dynamical system. This dynamical system consists of the Lebesgue measure on the unit interval with an incrementally constructed partition and transformation. In contrast, the RHA process method begins with some nonstationary nonergodic process from which we obtain a given stationary ergodic measure by taking the stationary mean and ergodic decomposition. For our particular application of constructing regular Hilberg processes, the RHA process method is sufficient and seems natural enough but it is likely insufficient for constructing processes which satisfy condition (5) without condition (6). In the later case, being the case of interest for modeling natural language, using the cutting and stacking method is a certain idea but we have not figured out yet how to implement it exactly.

To briefly explain our method, the RHA processes are formed in two not so complicated steps. First, we sample recursively random pools of k_n distinct blocks of length 2^n, which are formed by concatenation of randomly selected k_n pairs chosen from k_{n-1} distinct blocks of length 2^{n-1} sampled in a previous step (the recursion stops at blocks of length 1, which are fixed symbols). Second, we obtain an infinite sequence of random symbols by concatenating blocks of lengths 2^0, 2^1, 2^2, ... randomly chosen from the respective pools. As a result there cannot be more that k_n^2 distinct blocks of length 2^n that appear the final process realization. The selection of these blocks is, however, random and we do not know them a priori. This is some reason why the constructed process satisfies conditions similar to (5) and (6) simultaneously but is nonergodic.

Now we will write down this construction using symbols.

Step 1: Formally, let perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ be some sequence of strictly positive natural numbers that satisfy

$$k_{n-1} \le k_n \le k_{n-1}^2. \tag{9}$$

Next, for each $n \in \mathbb{N}$, let $(L_{nj}, R_{nj})_{j\in\{1,...,k_n\}}$ be an independent random combination of k_n pairs of numbers from the set $\{1, ..., k_{n-1}\}$ drawn without repetition. That is, we assume that each pair (L_{nj}, R_{nj}) is different, the elements of pairs may be identical ($L_{nj} = R_{nj}$), and the sequence $(L_{nj}, R_{nj})_{j\in\{1,...,k_n\}}$ is sorted lexicographically. Formally, we assume that random variables L_{nj} and R_{nj} are supported on some probability space $(\Omega, \mathcal{J}, P)$ and have the uniform distribution

$$\begin{aligned} P((L_{n1}, R_{n1}, ..., L_{nk_n}, R_{nk_n}) = (l_{n1}, r_{n1}, ..., l_{nk_n}, r_{nk_n})) \\ = \binom{k_{n-1}^2}{k_n}^{-1}. \end{aligned} \tag{10}$$

Subsequently we define random variables

$$Y_j^0 = j, \qquad j \in \{1, ..., k_0\}, \tag{11}$$

$$Y_j^n = Y_{L_{nj}}^{n-1} \times Y_{R_{nj}}^{n-1}, \qquad j \in \{1, ..., k_n\}, n \in \mathbb{N}, \tag{12}$$

where $a \times b$ denotes concatenation. Hence Y_j^n are k_n distinct blocks of 2^n natural numbers, selected by some sort of random hierarchical concatenation.

Step 2: Variables Y_j^n will be the building blocks of yet another process. Let $(C_n)_{n\in\{0\}\cup\mathbb{N}}$ be independent random variables, independent from $(L_{nj}, R_{nj})_{n\in\mathbb{N}, j\in\{1,...,k_n\}}$, with uniform distribution

$$P(C_n = j) = 1/k_n, \qquad j \in \{1, ..., k_n\}. \tag{13}$$

Definition 2 *The* random hierarchical association (RHA) process $\mathcal{X}$ with perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ *is defined as*

$$\mathcal{X} = Y_{C_0}^0 \times Y_{C_1}^1 \times Y_{C_2}^2 \times \ldots . \tag{14}$$

This completes the construction of the RHA processes but it is not the end of our discussion of these processes.

It is convenient to define a few more random variables for the RHA process. First, sequence $\mathcal{X}$ will be parsed into a sequence of numbers X_j, where

$$\mathcal{X} = X_1 \times X_2 \times X_3 \times ..., \tag{15}$$

and, second, we denote blocks starting at any position as

$$X_{k:l} = X_k \times X_{k+1} \times ... \times X_l. \tag{16}$$

The RHA processes defined in Definition 2 are not stationary but they possess a stationary mean, which is a condition related to asymptotic mean stationarity. Let us introduce shift operation $T : \mathbb{A}^{\mathbb{N}} \ni (x_i)_{i\in\mathbb{N}} \mapsto (x_{i+1})_{i\in\mathbb{N}} \in \mathbb{A}^{\mathbb{N}}$. We recall this definition:

Definition 3 *A measure ν on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ is called* asymptotically mean stationary (AMS) *if limits*

$$\mu(A) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nu(T^{-i}A) \tag{17}$$

exist for every event $A \in \mathcal{A}^{\mathbb{N}}$ [25].

For an AMS measure ν, function μ is a stationary measure on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$, called the stationary mean of ν. Moreover, measures μ and ν are equal on the shift invariant algebra $\mathcal{I} = \left\{A \in \mathcal{A}^{\mathbb{N}} : T^{-1}A = A\right\}$, i.e., $\mu(A) = \nu(A)$ for all $A \in \mathcal{I}$.

Now, let $\mathbb{A}^+ = \bigcup_{n\in\mathbb{N}} \mathbb{A}^n$. There is a related relaxed condition of asymptotic mean stationarity:

Definition 4 *A measure ν on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ is called* pseudo-asymptotically mean stationary (pseudo-AMS) *if limits*

$$\mu(x_{1:m}) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nu(\xi_{i:i+m-1} = x_{1:m}) \tag{18}$$

exist for every block $x_{1:m} \in \mathbb{A}^+$.

For a pseudo-AMS measure ν over a finite alphabet $\mathbb{A}$, function μ, extended via $\mu(\xi_{1:m} = x_{1:m}) := \mu(x_{1:m})$, is also a stationary measure on $(\mathbb{A}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$. We shall continue to call this μ a stationary mean of ν. However, a pseudo-AMS measure need not be AMS in general, cf. [26, Remark in the proof of Lemma 7.16] and [27, Example 6.3]. In particular, for a pseudo-AMS measure ν we need not have $\mu(A) = \nu(A)$ for shift invariant events $A \in \mathcal{I}$.

It turns out that the RHA processes are pseudo-AMS.

Theorem 4 *The RHA processes are pseudo-AMS. In particular, for $m \le 2^n$ and $k \in \mathbb{N}$, the stationary mean is*

$$\mu(x_{1:m}) = \frac{1}{2^n} \sum_{j=0}^{2^n-1} P(X_{k2^n+j:k2^n+j+m-1} = x_{1:m}). \tag{19}$$

The proof of Theorem 4 will be presented later in this article.

We suppose that the RHA processes are also AMS but we could not prove it so far. However, we have been able to show that certain RHA processes give rise to regular Hilberg processes:

Theorem 5 *For perplexities*

$$k_n = \left\lfloor \exp\left(2^{\beta n}\right) \right\rfloor, \tag{20}$$

where $0 < \beta < 1$, the stationary mean μ of the RHA process satisfies the following conditions:

(i) The Shannon entropy rate is $h_\mu = 0$.

(ii) The Shannon block entropy is sandwiched by

$$\frac{C_1 m}{(\log m)^\alpha} \le H_\mu(m) \le C_2 m \left(\frac{\log\log m}{\log m}\right)^\alpha, \tag{21}$$

where $\alpha = 1/\beta - 1$.

(iii) The stationary mean μ is a regular Hilberg process with exponent β.

(iv) The stationary mean μ is nonergodic and the Shannon entropy of the shift invariant algebra $H_\mu(\mathcal{I})$, as defined in [23], is infinite.

The proof of Theorem 5, which we consider the main result of this paper, will be postponed, as well. Although claim (i) follows from claim (iii) by Theorem 2, it will be established using a different method, of an independent interest.

Theorem 5 has some implications for universal coding. For a uniquely decodable code C, we denote its length for block $\xi_{1:m}$ as $|C(\xi_{1:m})|$. We recall that $\mathbf{E}_\mu |C(\xi_{1:m})| \geq H_\mu(m)$, so the Shannon block entropy provides a lower bound for compression of a stochastic process. In contrast, a code C is called universal if

$$\lim_{m\to\infty} \frac{|C(\xi_{1:m})|}{m} = h_\mu \tag{22}$$

holds almost surely for every stationary ergodic measure μ. Universal codes exist and the Lempel-Ziv code [28] is some example of such a code. The convergence rate for universal codes can be arbitrarily slow, however. Shields [29] showed that for any uniquely decodable code C and any sublinear function $\rho(m) = o(m)$ there exists such an ergodic source μ that

$$\limsup_{m\to\infty} [\mathbf{E}_\mu |C(\xi_{1:m})| - H_\mu(m) - \rho(m)] > 0. \tag{23}$$

Whereas Shields' result concerns nonexistence of a universal sublinear bound for the difference $|C(\xi_{1:m})| - H_\mu(m)$, some way of supplementing it is to investigate ratio $|C(\xi_{1:m})|/H_\mu(m)$. Although this ratio is asymptotically equal to 1 for universal codes and processes with a positive Shannon entropy rate $h_\mu > 0$, Shields' result does not predict how the ratio behaves for processes with a vanishing Shannon entropy rate $h_\mu = 0$. In fact, for the Lempel-Ziv code and ergodic regular Hilberg processes, there is no essentially sublinear bound for the ratio $|C(\xi_{1:m})|/H_\mu(m)$:

Theorem 6 *Let C be the Lempel-Ziv code. For an ergodic regular Hilberg process μ with exponent β, μ-almost surely*

$$\frac{|C(\xi_{1:m})|}{H_\mu(m)} = \Omega\left(\frac{m^{1-\beta}}{(\log m)^{1/\beta-1}}\right). \tag{24}$$

Proof: By ergodicity, we have $\mu = F$. Thus, by (7) and (6), we obtain

$$H_\mu(m) = H_F(m) \leq H_{top}(m|\xi_{1:\infty}) = O\left(m^\beta\right). \tag{25}$$

On the other hand, the length of the Lempel-Ziv code $|C(\xi_{1:m})|$ for a block $\xi_{1:m}$, by (5), μ-almost surely satisfies

$$\begin{aligned} |C(\xi_{1:m})| &\geq \frac{m}{L(\xi_{1:m})+1} \log \frac{m}{L(\xi_{1:m})+1} \\ &= \Omega\left(\frac{m}{(\log m)^{1/\beta-1}}\right). \end{aligned} \tag{26}$$

The first inequality in (26) stems from a simple observation in [11] that the length of the Lempel-Ziv code is greater than $V \log V$, where V is the number of Lempel-Ziv phrases, whereas the Lempel-Ziv phrases may not be longer than the maximal repetition plus 1. □

A somewhat more general result holds for the RHA processes from Theorem 5. In this case, we may replace the Lempel-Ziv code with an arbitrary uniquely decodable code:

Theorem 7 *Let C be an arbitrary uniquely decodable code. For the stationary mean μ of the RHA process with perplexities (20) and its random ergodic measure $F = \mu(\cdot|\mathcal{I})$, we have*

$$\mathbf{E}_\mu \frac{\mathbf{E}_F |C(\xi_{1:m})|}{H_F(m)} = \Omega\left(\frac{m^{1-\beta}}{(\log m)^{1/\beta-1}}\right), \tag{27}$$

Ratio (27) can be larger than any function $o(m^{1-\epsilon})$.

Proof: The claim follows by (7), (6), (21), and the source coding inequality

$$\mathbf{E}_\mu \mathbf{E}_F |C(\xi_{1:m})| = \mathbf{E}_\mu |C(\xi_{1:m})| \geq H_\mu(m). \tag{28}$$

□

Theorems 6 and 7 should be read as a warning that the length of a universal code $|C(\xi_{1:m})|$ is not a very reliable estimate of the Shannon block entropy $H_\mu(m)$ for an ergodic regular Hilberg process. Whereas, using a universal code, we can reliably estimate the Shannon entropy rate h_μ, the code length $|C(\xi_{1:m})|$ can be orders of magnitude larger than the Shannon block entropy $H_\mu(m)$.

The remaining parts of this article are devoted to proving the more involved Theorems 4 and 5. The organization is as follows. In Section II, some auxiliary notations are introduced. In Section III, Theorem 4 is demonstrated. In Section IV, the entropies and the maximal repetition for the RHA process and its stationary mean are related. Section V concerns some further auxiliary results, such as probabilities of no repeat and a bound for the topological entropy. In Section VI, Shannon block entropies of the RHA processes are discussed. In Section VII, Theorem 5 is proved.

II Auxiliary notations

Let us recall the construction of the RHA process from the previous section. In this section we introduce a few notations which will be used further. The collection of random variables (L_{nj}, R_{nj}) will be denoted as

$$\mathcal{G} = (L_{nj}, R_{nj})_{n\in\mathbb{N}, j\in\{1,...,k_n\}}. \tag{29}$$

We will also use notations

$$\mathcal{G}_{\leq m} = (L_{nj}, R_{nj})_{n\leq m, j\in\{1,...,k_n\}}, \tag{30}$$

$$\mathcal{G}_{>m} = (L_{nj}, R_{nj})_{n>m, j\in\{1,...,k_n\}}. \tag{31}$$

Let us observe that collection $\mathcal{G}_{\leq m}$ fully determines variables Y_j^m for a fixed m.

It is convenient to define a few more random variables for the RHA process. First, generalizing parsing (15), sequence $\mathcal{X}$ will be parsed into a sequence of blocks X_j^n of length 2^n, where

$$\mathcal{X} = Y_{C_0}^1 \times Y_{C_1}^1 \times Y_{C_2}^2 \times ... \times Y_{C_n}^n (= X_1^n) \times X_2^n \times X_3^n \times ... . \tag{32}$$

Let us also observe that there exist unique random variables K_{nj} such that

$$X_j^n = Y_{K_{nj}}^n. \tag{33}$$

Moreover, generalizing notation (16), we also denote blocks of length 2^n starting at any position as

$$X_{k:l}^n = X_k^n \times X_{k+1}^n \times ... \times X_l^n. \tag{34}$$

III Stationary mean

In this section, we will demonstrate Theorem 4. This theorem states that the RHA process has a stationary mean in a weaker sense, i.e., it is pseudo-asymptotically mean stationary (pseudo-AMS).

First we will prove this useful and a bit surprising property, which will be used in the present and in the further sections.

Proposition 1 *Variables K_{nj} are independent from $\mathcal{G}_{\le n}$ and satisfy*

$$P(K_{nj} = l, K_{n,j+1} = m) = 1/k_n^2, \qquad l, m \in \{1, ..., k_n\}, j \in \mathbb{N}. \tag{35}$$

Proof: Each K_{nj} is a function of C_q for some $q \ge n$ and $\mathcal{G}_{>n}$. Hence K_{nj} are independent from $\mathcal{G}_{\le n}$.

Now we will show by induction on j that (35) is satisfied.

The induction begins with $K_{n1} = C_n$ and $K_{n2} = L_{n+1,C_{n+1}}$. These two variables are independent by definition and by definition K_{n1} is uniformly distributed on $\{1, ..., k_n\}$. It remains to show that so is K_{n2}. Observe that $(L_{n+1,k}, R_{n+1,k})$ are independent of C_{n+1}. Hence for $l, m \in \{1, ..., k_n\}$ we obtain

$$\begin{aligned} P(K_{n2} = l, K_{n3} = m) &= \sum_{k=1}^{k_{n+1}} P(L_{n+1,k} = l, R_{n+1,k} = m) P(C_{n+1} = k) \\ &= \frac{1}{k_{n+1}} \sum_{k=1}^{k_{n+1}} P(L_{n+1,k} = l, R_{n+1,k} = m) \\ &= \frac{1}{k_{n+1}} \binom{k_n^2}{k_{n+1}}^{-1} \binom{k_n^2 - 1}{k_{n+1} - 1} = \frac{1}{k_{n+1}} \frac{k_{n+1}}{k_n^2} = \frac{1}{k_n^2}, \end{aligned}$$

so K_{n2} is uniformly distributed on $\{1, ..., k_n\}$.

The inductive step is as follows: (i) if $K_{n+1,j}$ is uniformly distributed on $\{1, ..., k_{n+1}\}$ then $(K_{n,2j}, K_{n,2j+1}) = (L_{n+1,K_{n+1,j}}, R_{n+1,K_{n+1,j}})$ is uniformly distributed on $\{1, ..., k_n\} \times \{1, ..., k_n\}$, and (ii) if $(K_{n+1,j}, K_{n+1,j+1})$ is uniformly distributed on $\{1, ..., k_{n+1}\} \times \{1, ..., k_{n+1}\}$ then $(K_{n,2j+1}, K_{n,2j+2}) = (R_{n+1,K_{n+1,j}}, L_{n+1,K_{n+1,j+1}})$ is uniformly distributed on $\{1, ..., k_n\} \times \{1, ..., k_n\}$. Now observe that $(L_{n+1,k}, R_{n+1,k})$ are independent of $K_{n+1,j}$. Hence, for

$l, m \in \{1, ..., k_n\}$ we obtain

$$\begin{aligned}
&P(K_{n,2j} = l, K_{n,2j+1} = m) \\
&= \sum_{k=1}^{k_{n+1}} P(L_{n+1,k} = l, R_{n+1,k} = m) P(K_{n+1,j} = k) \\
&= \frac{1}{k_{n+1}} \sum_{k=1}^{k_{n+1}} P(L_{n+1,k} = l, R_{n+1,k} = m) \\
&= \frac{1}{k_{n+1}} \binom{k_n^2}{k_{n+1}}^{-1} \binom{k_n^2 - 1}{k_{n+1} - 1} = \frac{1}{k_{n+1}} \frac{k_{n+1}}{k_n^2} = \frac{1}{k_n^2},
\end{aligned}$$

which proves claim (i). On the other hand, for $l, m \in \{1, ..., k_n\}$ we obtain

$$\begin{aligned}
&P(K_{n,2j+1} = l, K_{n,2j+2} = m) \\
&= \sum_{p,q=1}^{k_{n+1}} P(R_{n+1,p} = l, L_{n+1,q} = m) P(K_{n+1,j} = p, K_{n+1,j+1} = q) \\
&= \frac{1}{k_{n+1}^2} \sum_{p,q=1}^{k_{n+1}} P(R_{n+1,p} = l, L_{n+1,q} = m) \\
&= \frac{1}{k_{n+1}^2} \sum_{p=1}^{k_{n+1}} P(R_{n+1,p} = l, L_{n+1,p} = m) \\
&\quad + \frac{1}{k_{n+1}^2} \sum_{p,q=1,\, p \neq q}^{k_{n+1}} P(R_{n+1,p} = l, L_{n+1,q} = m) \\
&= \frac{1}{k_{n+1}^2} \binom{k_n^2}{k_{n+1}}^{-1} \left(\binom{k_n^2 - 1}{k_{n+1} - 1} + (k_n^2 - 1) \binom{k_n^2 - 2}{k_{n+1} - 2} \right) \\
&= \frac{1}{k_{n+1}^2} \left(\frac{k_{n+1}}{k_n^2} + (k_n^2 - 1) \frac{k_{n+1}(k_{n+1} - 1)}{k_n^2 (k_n^2 - 1)} \right) = \frac{1}{k_n^2},
\end{aligned}$$

which proves claim (ii). □

Using Proposition 1, it is easy to demonstrate Theorem 4.

Proof of Theorem 4: Block $X_{k2^n+j:k2^n+j+m-1}$ is a subsequence of $X^n_{k:k+1}$ for $m \le 2^n$, $k \in \mathbb{N}$, and $0 \le j < 2^n$. In particular, there exist functions f_{mj} such that

$$X_{k2^n+j:k2^n+j+m-1} = f_{mj}(X^n_{k:k+1}).$$

Hence probabilities $P(X_{i:i+m-1} = x_{1:m})$ are periodic functions of i with period 2^n, by Proposition 1. This implies the formula for $\mu(x_{1:m})$. □

IV Bounds for the stationary mean

This sections opens the discussion of various auxiliary results necessary to establish Theorem 5, the main result of this paper. The theorem operates with three functions of the stationary mean of the RHA process: Shannon block entropy, maximal repetition, and topological entropy. We first observe that it may be easier to analyze the behavior of blocks X^n_j drawn from the original the RHA process than the behavior of its stationary mean. For this reason, in this section we want to derive some bounds for the entropies and the maximal repetition of the stationary mean from the analogical bounds for blocks X^n_j. In the following we will denote

$$X^n_{kj} = X_{k2^n+j:k2^n+j+2^n-1}. \tag{36}$$

In particular, we have $X^n_{k0} = X^n_k$.

Subsequently, for Shannon entropy $H(X) = \mathbf{E}_P \left[-\log P(X)\right]$, we obtain:

Proposition 2 *For the stationary mean μ of the RHA process, we have*

$$H(X^{n-1}_j) \le H_\mu(2^n) \le H(X^{n+1}_j) + n \log 2. \tag{37}$$

Proof: By the Jensen inequality for function $p \mapsto -p \log p$ and Theorem 4, we hence obtain

$$H_\mu(2^n) \ge \frac{1}{2^n} \sum_{j=0}^{2^n-1} H(X^n_{kj}). \tag{38}$$

Now we observe that for each $k \ge 1$ and j there exists a q such that X^{n-1}_q is a subsequence of X^n_{kj}. Thus we have $H(X^n_{kj}) \ge H(X^{n-1}_q)$. This combined with inequality (38) yields $H(X^{n-1}_j) \le H_\mu(2^n)$. On the other hand, using inequality $\mu(x_{1:2^n}) \ge 2^{-n} P(X^n_{kj} = x_{1:2^n})$ and Theorem 4, we obtain

$$H_\mu(2^n) \le \frac{1}{2^n} \sum_{j=0}^{2^n-1} H(X^n_{kj}) + n \log 2. \tag{39}$$

Now we observe that for each $k > 1$ and j there exists a q such that X^n_{kj} is a subsequence of X^{n+1}_q. Thus we have $H(X^n_{kj}) \le H(X^{n+1}_q)$. This combined with inequality (39) yields $H_\mu(2^n) \le H(X^{n+1}_j) + n \log 2$. □

Analogically, we can bound the maximal repetition of the stationary mean. The result will be stated more generally. We will say that a function $\phi : \mathbb{A}^+ \to \mathbb{R}$ is increasing if for u being a subsequence of w, we have $\phi(u) \le \phi(w)$. Examples of increasing functions include the maximal repetition $L(w)$, the topological entropy $H_{top}(m|w)$, and the indicator function $\mathbf{1}\{\phi(w) > k\}$, where ϕ is increasing.

Proposition 3 *For the stationary mean μ of the RHA process and an increasing function ϕ, we have*

$$\mathbf{E}_P\, \phi(X^{n-1}_j) \le \mathbf{E}_\mu\, \phi(\xi_{1:2^n}) \le \mathbf{E}_P\, \phi(X^{n+1}_j). \tag{40}$$

Proof: By Theorem 4,

$$\mathbf{E}_\mu\, \phi(\xi_{1:2^n}) = \frac{1}{2^n} \sum_{j=0}^{2^n-1} \mathbf{E}_P\, \phi(X^n_{kj}). \tag{41}$$

Now we observe that for each $k \geq 1$ and j there exists a q such that X_q^{n-1} is a subsequence of X_{kj}^n. Thus we have $\phi(X_{kj}^n) \geq \phi(X_q^{n-1})$. This combined with equality (41) yields $\mathbf{E}_P\, \phi(X_j^{n-1}) \leq \mathbf{E}_\mu\, \phi(\xi_{1:2^n})$. On the other hand, for each $k > 1$ and j there exists a q such that X_{kj}^n is a subsequence of X_q^{n+1}. Thus we have $\phi(X_{kj}^n) \leq \phi(X_q^{n+1})$. This combined with equality (41) yields $\mathbf{E}_\mu\, \phi(\xi_{1:2^n}) \leq \mathbf{E}_P\, \phi(X_j^{n+1})$. □

Hence, to obtain the desired bounds for the stationary mean, it suffices to investigate the distribution of blocks X_j^n.

V Further auxiliary results

To make another observation, Theorem 5 links the Shannon block entropy, maximal repetition and topological entropy of the RHA process with its parameters called perplexities k_n. Therefore, the goal of this section is to furnish some bounds for topological entropy and maximal repetition of blocks X_{kj}^n in terms of perplexities k_n. In contrast, in the next section we will use perplexities k_n to bound the Shannon entropies of blocks X_{kj}^n.

Let us begin with a simple lower bound for the topological entropy of blocks X_j^n. From this bound we can then obtain an upper bound for the maximal repetition by Theorem 1.

Proposition 4 *For the RHA process, almost surely*

$$H_{top}(2^m|\mathcal{X}) \leq 2\log k_m. \tag{42}$$

Proof: For a given realization of the RHA process (i.e., for fixed Y_j^m), there are at most k_m different values of blocks X_j^m. Therefore, there are at most k_m^2 different values of blocks X_{kj}^m in sequence $\mathcal{X}$. □

Obtaining a lower bound for the topological entropy and an upper bound for the maximal repetition of blocks X_j^n is more involved. These topics will be discussed in the following sections. For this goal, we will consider events $A_{n,-1} := \emptyset$ and

$$A_{nm} := (X_1^n \text{ consists of } 2^{n-m} \text{ distinct blocks } X_j^m) \tag{43}$$

We have $P(A_{nn}) = 1$ and $A_{nm} \supset A_{n,m-1}$. Probabilities $P(A_{nm})$ will be called probabilities of no repeat.

Proposition 5 *For the RHA process, we have $P(A_{nm}) = 0$ for $k_m < 2^{n-m}$, whereas for $k_m \geq 2^{n-m}$ and $m < n$ we have*

$$P(A_{nm}) = P(A_{n,m+1})\frac{k_m(k_m-1)\ldots(k_m-2^{n-m}+1)}{k_m^2(k_m^2-1)\ldots(k_m^2-2^{n-m-1}+1)}. \tag{44}$$

Proof: There are no more than k_m distinct blocks X_j^m in block X_1^n. Thus $P(A_{nm}) = 0$ for $k_m < 2^{n-m}$. Now assume $k_m \geq 2^{n-m}$. Introduce random variables D_{mi} such that $X_1^n = Y_{D_{m1}}^m \times ... \times Y_{D_{m2^{n-m}}}^m$. Consider probabilities

$p_m = P(D_{m1} = d_1, ..., D_{m2^{n-m}} = d_{2^{n-m}})$, where d_i are distinct. It can be easily shown by induction on decreasing m that p_m do not depend on d_i and satisfy

$$p_m = p_{m+1} \binom{k_m^2}{k_{m+1}}^{-1} \binom{k_m^2 - 2^{n-m-1}}{k_{m+1} - 2^{n-m-1}}.$$

Moreover, since p_m do not depend on d_i, we obtain $P(A_{nm}) = p_m k_m (k_m - 1) \dots (k_m - 2^{n-m} + 1)$. Hence the claim follows. □

VI Shannon block entropy

This section is the last preparatory section. Here we will bound the Shannon entropies of blocks X_j^n in terms of perplexities k_n. To establish some necessary notation, for random variables X, Y and Z, where X is discrete whereas Y and Z need not be so, besides Shannon entropy $H(X) = \mathbf{E}_P [- \log P(X)]$, we define conditional entropy $H(X|Y) = \mathbf{E}_P [- \log P(X|Y)]$, mutual information $I(X;Y) := H(X) - H(X|Y)$, and conditional mutual information $I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$. Given these objects, we will bound the Shannon entropies of blocks of the RHA process.

The first result is a corollary of Proposition 1, which says that conditional entropy of blocks X_j^n given the entire pool of admissible blocks of the same length $\mathcal{G}_{\le n}$ is exactly equal to the logarithm of perplexity.

Proposition 6 *We have*

$$H(X_j^n | \mathcal{G}_{\le n}) = \log k_n \tag{45}$$

and $I(X_j^n; X_{j+1}^n | \mathcal{G}_{\le n}) = 0$.

Proof: Given $\mathcal{G}_{\le n}$, the correspondence between X_j^n and K_{nj} is one-to-one. Hence $H(X_j^n | \mathcal{G}_{\le n}) = H(K_{nj} | \mathcal{G}_{\le n})$. From Proposition 1 we further obtain $H(K_{nj} | \mathcal{G}_{\le n}) = H(K_{nj}) = \log k_n$ and $H(K_{nj}, K_{n,j+1} | \mathcal{G}_{\le n}) = H(K_{nj}) + H(K_{n,j+1})$. □

The second result is an exact expression for the Shannon entropy of the pool of admissible blocks $\mathcal{G}_{\le n}$, also in term of perplexities.

Proposition 7 *We have*

$$H(\mathcal{G}_{\le n}) = \sum_{l=1}^{n} \log \binom{k_{l-1}^2}{k_l}. \tag{46}$$

Proof: The claim follows by chain rule $H(\mathcal{G}_{\le n}) = H(\mathcal{G}_{\le n-1}) + H(\mathcal{G}_{\le n} | \mathcal{G}_{\le n-1})$ from $H(\mathcal{G}_{\le 0}) = 0$ and $H(\mathcal{G}_{\le n} | \mathcal{G}_{\le n-1}) = \log \binom{k_{n-1}^2}{k_n}$. □

Combining the above two results, we can provide an upper bound for the unconditional Shannon entropy of blocks X_j^n.

Proposition 8 *We have*

$$H(X_j^n) \le \min_{0 \le l \le n} \left(H(\mathcal{G}_{\le l}) + 2^{n-l} \log k_l \right). \tag{47}$$

Proof: For any $0 \le l \le n$ we have $H(X_j^n) \le H(X_j^n, \mathcal{G}_{\le l}) = H(X_j^n|\mathcal{G}_{\le l}) + H(\mathcal{G}_{\le l})$, whereas $H(X_j^n|\mathcal{G}_{\le l}) \le 2^{n-l} H(K_{lj}|\mathcal{G}_{\le l}) = 2^{n-l} H(K_{lj}) = 2^{n-l} \log k_l$. □

Given Propositions 6 and 8, we may introduce an important parameter of the RHA process, which we will call the combinatorial entropy rate.

Definition 5 *The combinatorial entropy rate of the RHA process is*

$$h := \inf_{l \in \mathbb{N}} 2^{-l} \log k_l = \lim_{l \to \infty} 2^{-l} \log k_l. \tag{48}$$

Proposition 9 *We have*

$$\inf_{n \in \mathbb{N}} 2^{-n} H(X_j^n) = h. \tag{49}$$

Proof: On the one hand, by Proposition 6,

$$\inf_{n \in \mathbb{N}} 2^{-n} H(X_j^n) \ge \inf_{n \in \mathbb{N}} 2^{-n} H(X_j^n|\mathcal{G}_{\le n}) = \inf_{l \in \mathbb{N}} 2^{-l} \log k_l.$$

On the other hand, by Proposition 8,

$$\inf_{n \in \mathbb{N}} 2^{-n} H(X_j^n) \le \inf_{l \in \mathbb{N}} \inf_{n \in \mathbb{N}} \left(2^{-n} H(\mathcal{G}_{\le l}) + 2^{-l} \log k_l \right) = \inf_{l \in \mathbb{N}} 2^{-l} \log k_l.$$

□

Proposition 9 combined with Proposition 2 yields a bound for the Shannon entropy rate of the stationary mean of the RHA process.

Proposition 10 *For the stationary mean μ of the RHA process, we have*

$$h/2 \le h_\mu \le 2h. \tag{50}$$

Proof: Divide inequality (37) by 2^n and take the infimum. □

In particular, the combinatorial entropy rate vanishes ($h = 0$) if and only if the Shannon entropy rate of the stationary mean vanishes ($h_\mu = 0$) as well. This happens in particular for perplexities (20).

Inequality $H(X_j^n) \ge H(X_j^n|\mathcal{G}_{\le n}) = \log k_n$ gives a certain lower bound for the Shannon block entropy of the RHA process. For perplexities (20), this lower bound is orders of magnitude smaller than the upper bound (47). Concluding this section we would like to produce a lower bound which is of comparable order to (47).

Proposition 11 *We have*

$$H(X_j^n) \ge \max_{0 \le l \le n} \left(\log \binom{k_{l-1}^2}{k_l} - \log \binom{k_{l-1}^2 - 2^{n-l}}{k_l - 2^{n-l}} \right) P(A_{nl}), \tag{51}$$

where $P(A_{nl})$ are the probabilities of no repeat (44).

Proof: We have

$$H(X_j^n) \ge I(X_j^n; \mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}) = H(\mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}) - H(\mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}, X_j^n).$$

We have $H(\mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}) = \log\binom{k_{l-1}^2}{k_l}$. As for $H(\mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}, X_j^n)$, we may propose the following bound. Given X_j^n consisting of 2^{n-l} distinct blocks of length 2^l, tuple $(L_{lj}, R_{lj})_{j\in\{1,\dots,k_l\}}$ may assume at most $\binom{k_{l-1}^2 - 2^{n-l}}{k_l - 2^{n-l}}$ distinct values. Hence

$$H(\mathcal{G}_{\le l}|\mathcal{G}_{\le l-1}, X_j^n) \le P(A_{nl}) \log\binom{k_{l-1}^2 - 2^{n-l}}{k_l - 2^{n-l}},$$

from which the claim follows. □

VII Main result

Now we can demonstrate the main result, which will conclude our paper.

Proof of Theorem 5:

(i) For perplexities (20) the combinatorial entropy rate is $h = 0$. Hence $h_\mu = 0$ by Proposition 10.

(ii) By (46), entropy $H(\mathcal{G}_{\le n})$ can be bounded as

$$H(\mathcal{G}_{\le n}) = \sum_{l=1}^{n} \log\binom{k_{l-1}^2}{k_l} \le \sum_{l=1}^{n} 2k_l \log k_{l-1} \le 2nk_n \log k_n.$$

Hence, from (47), for $0 \le l \le n$ we obtain an upper bound:

$$H(X_j^n) \le \left(2lk_l + 2^{n-l}\right) \log k_l.$$

If we choose $l = \left\lfloor \beta^{-1} \log_2\left(\frac{n\log 2}{\log n}\right)\right\rfloor$ then for perplexities (20) we obtain

$$\begin{aligned} H(X_j^n) &\le \left[2\beta^{-1} \log_2\left(\frac{n\log 2}{\log n}\right) 2^{n/\log n} + 2^n \left(\frac{n\log 2}{\log n}\right)^{-1/\beta}\right] \frac{n\log 2}{\log n} \\ &= \Theta\left(2^n \left(\frac{\log n}{n}\right)^{1/\beta - 1}\right). \end{aligned} \tag{52}$$

On the other hand, from (51) and (44), for $0 \le l \le n$ we have

$$\begin{aligned} H(X_j^n) &\ge \left(\log\binom{k_{l-1}^2}{k_l} - \log\binom{k_{l-1}^2 - 2^{n-l}}{k_l - 2^{n-l}}\right) P(A_{nl}) \\ &\ge 2^{n-l} \log\left(\frac{k_{l-1}^2 - 2^{n-l} + 1}{k_l - 2^{n-l} + 1}\right) P(A_{nl}), \end{aligned}$$

where

$$\begin{aligned} P(A_{nl}) &= \prod_{m=l}^{n-1} \frac{k_m(k_m-1)\dots(k_m-2^{n-m}+1)}{k_m^2(k_m^2-1)\dots(k_m^2-2^{n-m-1}+1)} \\ &\ge \prod_{m=l}^{n-1} \left(\frac{(k_m-2^{n-m}+2)(k_m-2^{n-m}+1)}{k_m^2-2^{n-m-1}+1}\right)^{2^{n-m-1}} \\ &\ge \left(\frac{(k_l-2^{n-l}+2)(k_l-2^{n-l}+1)}{k_l^2-2^{n-l-1}+1}\right)^{\sum_{m=l}^{n-1} 2^{n-m-1}} \\ &\ge \left(1-\frac{k_l(2^{n-l+1}-3)+2}{k_l^2-2^{n-l-1}+1}\right)^{2^n} \\ &\ge 1-2^n\frac{k_l(2^{n-l+1}-3)+2}{k_l^2-2^{n-l-1}+1}. \end{aligned} \tag{53}$$

If we choose $l = \lceil \beta^{-1}\log_2(2n)\rceil$ then for perplexities (20) we obtain that $k_l > \exp(2n) > 2^{2n}$. Hence $P(A_{nl})$ is greater than a certain constant $\alpha > 0$ and

$$H(X_j^n) \ge \alpha 2^n (2n)^{-1/\beta}[2^{1-\beta}-1]2n = \Theta\left(2^n\left(\frac{1}{n}\right)^{1/\beta-1}\right). \tag{54}$$

By (52) and (54), from Proposition 2, we obtain the desired sandwich bound for the entropy of the stationary mean.

(iii) By Proposition 4 and Proposition 3 we obtain

$$\begin{aligned} 0 &= \mathbf{E}_P\,\mathbf{1}\{H_{top}(2^m|\mathcal{X}) > 2\log k_m\} \\ &\ge \mathbf{E}_P\,\mathbf{1}\left\{H_{top}(2^m|X_j^{n+1}) > 2\log k_m\right\} \\ &\ge \mathbf{E}_\mu\,\mathbf{1}\{H_{top}(2^m|\xi_{1:2^n}) > 2\log k_m\}. \end{aligned}$$

Hence μ-almost surely $H_{top}(2^m|\xi_{1:\infty}) \le 2\log k_m = 2^{\beta m+1}$, which implies the upper bound $H_{top}(m|\xi_{1:\infty}) < C_1 m^\beta$ for a certain constant C_1. From this we obtain the lower bound $L(\xi_{1:m}) > C_2(\log m)^{1/\beta}$ by Theorem 1.

As for the converse bounds, we have $L(X_1^n) \ge 2^l$ for A_{nl}^c, where A_{nl} are the events of no repeat (43). Hence by Proposition 3,

$$\mathbf{E}_\mu\,\mathbf{1}\{L(\xi_{1:2^n}) \ge l\} \le \mathbf{E}_P\,\mathbf{1}\left\{L(X_l^{n+1}) \ge l\right\} \le 1 - P(A_{n+1,l}).$$

Now, if we choose $l = \lceil \beta^{-1}\log_2(2n)\rceil$ then for perplexities (20) we obtain that $k_l > \exp(2n) > 2^{2n}$. Hence, by (53), $\sum_{n=0}^\infty (1-P(A_{n+1,l})) < \infty$. Consequently, by the Borel-Cantelli lemma $L(\xi_{1:2^n}) < l$ must hold for sufficiently large n μ-almost surely. Thus $L(\xi_{1:m}) < C_3(\log m)^{1/\beta}$ for sufficiently large m. From this we obtain the lower bound $H_{top}(m|\xi_{1:\infty}) > C_4 m^\beta$ for sufficiently large m by Theorem 1.

(iv) Denote the random ergodic measure $F = \mu(\cdot|\mathcal{I})$ of the stationary mean μ. The entropy of the shift-invariant algebra with respect to μ may be bounded by mutual information as

$$\begin{aligned} H_\mu(\mathcal{I}) = \lim_{m\to\infty} I_\mu(\mathcal{I};\xi_{1:m}) &= \lim_{m\to\infty} \left[H_\mu(\xi_{1:m}) - H_\mu(\xi_{1:m}|\mathcal{I})\right] \\ &= \lim_{m\to\infty} \left[H_\mu(m) - \mathbf{E}_\mu\, H_F(m)\right] \\ &= \lim_{m\to\infty} \left[H_\mu(m) - \mathbf{E}_\mu\, H_{top}(m|\xi_{1:\infty})\right] = \infty. \end{aligned}$$

Since the entropy of the shift-invariant algebra is strictly positive, the measure μ is nonergodic.

□

Acknowledgment

The author wishes to thank Jan Mielniczuk and an anonymous referee for valuable comments and literature suggestions.

References

[1] A. de Luca, “On the combinatorics of finite words,” *Theor. Comput. Sci.*, vol. 218, pp. 13–39, 1999.

[2] P. C. Shields, “String matching: The ergodic case,” *Ann. Probab.*, vol. 20, pp. 1199–1203, 1992.

[3] ——, “String matching bounds via coding,” *Ann. Probab.*, vol. 25, pp. 329–336, 1997.

[4] R. Kolpakov and G. Kucherov, “Finding maximal repetitions in a word in linear time,” in *40th Annual Symposium on Foundations of Computer Science, 1999*, 1999, pp. 596–604.

[5] ——, “On maximal repetitions in words,” *J. Discr. Algor.*, vol. 1, pp. 159–186, 1999.

[6] S. Janson, S. Lonardi, and W. Szpankowski, “On average sequence complexity,” *Theor. Comput. Sci.*, vol. 326, pp. 213–227, 2004.

[7] S. Ferenczi, “Complexity of sequences and dynamical systems,” *Discr. Math.*, vol. 206, pp. 145–154, 1999.

[8] I. Gheorghiciuc and M. D. Ward, “On correlation polynomials and subword complexity,” *Discr. Math. Theo. Comp. Sci.*, vol. AH, pp. 1–18, 2007.

[9] E. E. Ivanko, “Exact approximation of average subword complexity of finite random words over finite alphabet,” *Trud. Inst. Mat. Meh. UrO RAN*, vol. 14, no. 4, pp. 185–189, 2008.

[10] Ł. Dębowski, “Estimation of entropy from subword complexity,” in *Challenges in Computational Statistics and Data Mining*, S. Matwin and J. Mielniczuk, Eds. Springer, 2016, pp. 53–70.

[11] ——, "Hilberg's conjecture — a challenge for machine learning," *Schedae Inform.*, vol. 23, pp. 33–44, 2014.

[12] W. Hilberg, "Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente?" *Frequenz*, vol. 44, pp. 243–248, 1990.

[13] Ł. Dębowski, "Maximal repetitions in written texts: Finite energy hypothesis vs. strong Hilberg conjecture," *Entropy*, vol. 17, pp. 5903–5919, 2015.

[14] ——, "Maximal lengths of repeat in English prose," in *Synergetic Linguistics. Text and Language as Dynamic System*, S. Naumann, P. Grzybek, R. Vulanović, and G. Altmann, Eds. Wien: Praesens Verlag, 2012, pp. 23–30.

[15] C. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, 1951.

[16] T. M. Cover and R. C. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inform. Theory*, vol. 24, pp. 413–421, 1978.

[17] P. F. Brown, S. D. Pietra, V. J. D. Pietra, J. C. Lai, and R. L. Mercer, "An estimate of an upper bound for the entropy of English," *Comput. Linguist.*, vol. 18, no. 1, pp. 31–40, 1983.

[18] P. Grassberger, "Data compression and entropy estimates by non-sequential recursive pair substitution," 2002, http://xxx.lanl.gov/abs/physics/0207023.

[19] F. Behr, V. Fossum, M. Mitzenmacher, and D. Xiao, "Estimating and comparing entropy across written natural languages using PPM compression," in *Proceedings of Data Compression Conference 2003*, 2003, p. 416.

[20] R. Takahira, K. Tanaka-Ishii, and Ł. Dębowski, "Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora," *Entropy*, vol. 18, no. 10, p. 364, 2016.

[21] O. Kallenberg, *Foundations of Modern Probability.* Springer, 1997.

[22] R. M. Gray and L. D. Davisson, "The ergodic decomposition of stationary discrete random processses," *IEEE Trans. Inform. Theory*, vol. 20, pp. 625–636, 1974.

[23] Ł. Dębowski, "A general definition of conditional information and its application to ergodic decomposition," *Statist. Probab. Lett.*, vol. 79, pp. 1260–1268, 2009.

[24] P. C. Shields, "Cutting and stacking: A method for constructing stationary processes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1605–1617, 1991.

[25] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, pp. 962–973, 1980.

[26] R. M. Gray, *Probability, Random Processes, and Ergodic Properties.* Springer, 2009.

[27] Ł. Dębowski, “Variable-length coding of two-sided asymptotically mean stationary measures,” *J. Theor. Probab.*, vol. 23, pp. 237–256, 2010.

[28] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Trans. Inform. Theory*, vol. 23, pp. 337–343, 1977.

[29] P. C. Shields, “Universal redundancy rates don’t exist,” *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 520–524, 1993.